



Methods and Techniques for Generation and Integration of Web Ontology Data

A Thesis Submitted for the Degree of
Doctor of Philosophy

By
Chao Wang

in

FACULTY OF INFORMATION TECHNOLOGY
UNIVERSITY OF TECHNOLOGY, SYDNEY
AUSTRALIA
JUNE 2007

UNIVERSITY OF TECHNOLOGY, SYDNEY
FACULTY OF INFORMATION TECHNOLOGY

The undersigned hereby certify that they have read this thesis entitled “**Methods and Techniques for Generation and Integration of Web Ontology Data**” by **Chao Wang** and that in their opinions it is fully adequate, in scope and in quality, as a thesis for the degree of **Doctor of Philosophy**.

Dated: June 2007

Research Supervisors: _____
Dr Jie Lu

Dr Guangquan Zhang

CERTIFICATE

Date: June 2007

Author: Chao Wang

Title: Methods and Techniques for Generation and
Integration of Web Ontology Data

Degree: Ph.D.

I certify that this thesis has not already been submitted for any degree and is not being submitted as part of candidature for any other degree.

I also certify that the thesis has been written by me and that any help that I have received in preparing this thesis, and all sources used, have been acknowledged in this thesis.



Signature of Author

Acknowledgements

I would like to express my sincere and deep gratitude to my principal supervisor, Jie Lu, for her continuous encouragement, advice, help and invaluable suggestions. Without her support, I couldn't even start the PhD study. She is such a nice, generous, helpful and kindhearted person. I feel really happy, comfortable and unconstrained with her during my PhD study. I owe my research achievements to her professional supervision.

Many thanks are also due to my co-supervisor, Guangquan Zhang, for his valued suggestions and constant support, and for the numerous conversations with him.

I wish to thank my fellow research students and the staff of the Faculty of Information Technology, University of Technology, Sydney (UTS). Their various assistance and advice are of great benefit to this study.

I appreciate the financial support from both the UTS International Research Scholarship (IRS) and the Faculty of Information Technology Scholarship. I appreciate the travel support for attending the international conferences which I received from the Faculty of Information Technology and the UTS Vice-Chancellor's Conference Fund.

I would like to express my heartfelt appreciation to my parents for their constant support, encouragement and care over all these years. Also, I would like to thank my wife Kei Shum for her support and love during my research in the past years.

To my parents, Limin Wang and Weijuan Feng

Table of Contents

| | |
|--------------------------------------------------------------------------------------------------------|-----------|
| Table of Contents | ix |
| List of Tables | x |
| List of Figures | xi |
| Abstract | xiii |
| 1 Introduction | 1 |
| 1.1 Research Issues and Motivations | 1 |
| 1.2 Contributions | 5 |
| 1.2.1 Key Information Mining Method and Its Application to Web Ontology Schema Generation | 5 |
| 1.2.2 Methods for Web Ontology Data Generation and Management | 6 |
| 1.2.3 Web Ontology Data Matching Methods for Integration | 7 |
| 1.3 Organization of the Thesis | 8 |
| 1.4 Publications Related to the Thesis | 11 |
| 2 Background and Literature Review | 13 |
| 2.1 Introduction | 13 |
| 2.2 Ontology | 13 |
| 2.2.1 Definition of Ontology | 14 |
| 2.2.2 Classification of Ontologies | 14 |
| 2.2.3 Ontology Languages | 16 |
| 2.3 Ontology Related Research Areas | 19 |
| 2.3.1 Semantic Web | 19 |
| 2.3.2 Web Intelligence | 21 |
| 2.3.3 E-Service Intelligence | 22 |
| 2.4 Ontology Generation | 24 |

| | | |
|----------|--------------------------------------------------------------------------|-----------|
| 2.4.1 | Ontology Schema and Data | 24 |
| 2.4.2 | Ontology Schema Generation | 25 |
| 2.4.3 | Ontology Data Generation | 29 |
| 2.5 | Ontology-based Data Integration and Matching | 32 |
| 2.5.1 | Data Semantics and Representations | 33 |
| 2.5.2 | Data Distribution and Management | 34 |
| 2.5.3 | Data Integration and Match Approaches | 35 |
| 2.6 | Summary | 39 |
| 3 | Ontology Schema Generation by Key Information Mining from the Web | 41 |
| 3.1 | Introduction | 41 |
| 3.2 | Key Information and Common Features of Web Pages | 43 |
| 3.3 | A Key Information Mining Method | 44 |
| 3.3.1 | Overview of the Method | 44 |
| 3.3.2 | Generation of Candidate Key Information | 45 |
| 3.3.3 | Entropy Evaluation | 49 |
| 3.4 | Experiments | 52 |
| 3.5 | Domain Ontology Schema Generation with KIM | 53 |
| 3.6 | Summary | 56 |
| 4 | Web Ontology Data Generation | 57 |
| 4.1 | Introduction | 57 |
| 4.2 | Web Ontology Data Conversion | 58 |
| 4.2.1 | Data Formats | 58 |
| 4.2.2 | Using XQuery for Conversion | 60 |
| 4.3 | Web Ontology Data Authoring | 61 |
| 4.3.1 | System Overview | 62 |
| 4.3.2 | Ontology Schema Browsing | 64 |
| 4.3.3 | Ontology Data Authoring | 67 |
| 4.4 | Discussion | 70 |
| 4.5 | Summary | 72 |
| 5 | Web Ontology Data Management | 73 |
| 5.1 | Introduction | 73 |
| 5.2 | A General Semantic Web Content Management Model | 75 |
| 5.2.1 | Semantic Wikis | 76 |
| 5.2.2 | Data focused Management | 76 |
| 5.3 | The Data focused SWCM Model | 77 |

| | | |
|----------|------------------------------------------------------------------|------------|
| 5.3.1 | Semantic Contents | 77 |
| 5.3.2 | Participants | 78 |
| 5.3.3 | Operations | 78 |
| 5.3.4 | Rules | 80 |
| 5.4 | System Design and Implementation Issues | 83 |
| 5.4.1 | System Overview | 83 |
| 5.4.2 | On Implementing Operations | 86 |
| 5.4.3 | On Implementing the Rule Engine | 87 |
| 5.5 | Further Discussion | 88 |
| 5.5.1 | Comparison with Other Systems | 88 |
| 5.5.2 | Towards Semantic Web Knowledge Management | 89 |
| 5.6 | Summary | 90 |
| 6 | Ontology Data Matching by Constrained Clustering | 92 |
| 6.1 | Introduction | 92 |
| 6.2 | Scenarios | 94 |
| 6.2.1 | A Scenario Related to the Semantic Web | 94 |
| 6.2.2 | More Scenarios in Other Areas | 96 |
| 6.3 | Characteristics of Ontology Data | 97 |
| 6.4 | Duplication in Ontology Data | 97 |
| 6.4.1 | Notations | 97 |
| 6.4.2 | Constraint Rules | 98 |
| 6.5 | A Constrained Clustering Approach | 100 |
| 6.5.1 | Canopy Clustering | 101 |
| 6.5.2 | Canopy Clustering with Constraints | 101 |
| 6.5.3 | Computational Complexity | 103 |
| 6.6 | Experiments | 105 |
| 6.7 | Summary | 108 |
| 7 | Ontology Data Matching through Learning Adaptive Metrics | 109 |
| 7.1 | Introduction | 109 |
| 7.2 | Similarity Metrics for Instance Matching | 111 |
| 7.2.1 | String Edit Distance and TF-IDF | 111 |
| 7.2.2 | Exploring Ontology Features for New Metrics | 112 |
| 7.3 | Adaptive Metrics by the Learning Mechanism | 114 |
| 7.4 | Experiments | 116 |
| 7.4.1 | Data Collection | 116 |
| 7.4.2 | Methodology and Results | 117 |
| 7.5 | A Peer-to-peer Framework for Ontology Data Integration | 121 |

| | | |
|----------|-----------------------------------------------------------|------------|
| 7.5.1 | Overview of the Framework | 121 |
| 7.5.2 | Ontology Data Matching in the Framework | 123 |
| 7.6 | Summary | 124 |
| 8 | Conclusions and Future Work | 126 |
| 8.1 | Conclusions | 126 |
| 8.2 | Future Work | 128 |
| 8.2.1 | More Functions and Applications with the KIM Method . . . | 128 |
| 8.2.2 | Improving Web Ontology Data Generation and Management . | 129 |
| 8.2.3 | Improving Ontology Data Matching Methods | 129 |
| 8.2.4 | Towards Applications in Real World Domains and Challenges | 130 |
| | Bibliography | 133 |

List of Tables

| | | |
|-----|-------------------------------------------------------------------------|-----|
| 2.1 | Semantic spectrum of ontologies | 15 |
| 2.2 | Different ontology learning approaches | 27 |
| 3.1 | Experiment results for five web sites | 53 |
| 6.1 | Performance of different approaches of duplicate detection | 107 |
| 6.2 | Precision of detection of duplicated pairs in different steps | 107 |
| 7.1 | Match distribution of instances in the ontology data set | 117 |
| 7.2 | Averaged Maximum F measure of different methods | 119 |

List of Figures

| | | |
|-----|----------------------------------------------------------------------------------------------------|----|
| 1.1 | The overall structure of the thesis | 10 |
| 2.1 | The ontology Spectrum according to the classification by McGuinness . | 15 |
| 2.2 | An example of a simple DL knowledge base consisting of a TBox and an ABox | 17 |
| 2.3 | Some examples of concepts, properties and individuals in OWL . . . | 18 |
| 2.4 | The layered architecture of semantic web proposed by Berners-Lee . . | 20 |
| 2.5 | Four levels of web intelligence proposed by Zhong | 21 |
| 3.1 | Word length in menus/navigation indicators of different web sites . . | 45 |
| 3.2 | Overview of the KIM method | 46 |
| 3.3 | Two illustrative menu subtrees | 48 |
| 3.4 | An algorithm for extracting a menu instance list from a web page . . | 49 |
| 3.5 | The interface of key information mining and OWL class marking & exportation | 54 |
| 3.6 | The exported raw owl file opened in Protégé for further editing and refinement | 55 |
| 4.1 | An XQuery used to convert XML into OWL | 61 |
| 4.2 | The general structure of the prototype system <i>robinet</i> | 63 |
| 4.3 | The interface that allows users to browse the structure of the domain ontology schema | 65 |

| | | |
|-----|--------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 4.4 | The interface that shows the details of the class <code>Article</code> in the ontology schema | 66 |
| 4.5 | The interface used for creating an instance of the class <code>Article</code> . . . | 68 |
| 4.6 | The created instance of the class <code>Article</code> | 69 |
| 4.7 | The instance after related instances are created and automatically linked with it | 71 |
| 5.1 | Fragments of an ontology schema reflecting the domain of company information management | 79 |
| 5.2 | The general structure of the extended <i>robinet</i> system that supports SWCM | 84 |
| 5.3 | A web page that shows the content of an instance of the class <code>Employee</code> | 85 |
| 5.4 | A web page showing that the value of <code>has-salary</code> can not be changed by the employee himself according to the imposed rules | 87 |
| 5.5 | The semantic web knowledge management over content management | 90 |
| 6.1 | An example of the ontology data in the semantic web scenario | 95 |
| 6.2 | An illustration of applications of different constraint rules in corresponding situations | 100 |
| 6.3 | The algorithm of Step I | 102 |
| 6.4 | The algorithm of Step II | 103 |
| 6.5 | The algorithm of Step III | 104 |
| 6.6 | The algorithm of Step IV | 105 |
| 6.7 | The sensitiveness of precision to T_{loose} for different approaches of duplicate detection | 108 |
| 7.1 | Recall/precision curves of different methods | 119 |
| 7.2 | F measure scores of different methods on two major classes | 120 |
| 7.3 | The framework that supports ontology generation and matching . . . | 122 |

Abstract

Data integration over the web or across organizations encounters several unfavorable features: heterogeneity, decentralization, incompleteness, and uncertainty, which prevent information from being fully utilized for advanced applications such as decision support services. The basic idea of ontology related approaches for data integration is to use one or more ontology schemas to interpret data from different sources. Several issues will come up when actually implementing the idea: (1) How to develop the domain ontology schema(s) used for the integration; (2) How to generate ontology data for domain ontology schema if the data are not in the right format and to create and manage ontology data in an appropriate way; (3) How to improve the quality of integrated ontology data by reducing duplications and increasing completeness and certainty.

This thesis focuses on the above issues and develops a set of methods to tackle them.

First, a key information mining method is developed to facilitate the development of interested domain ontology schemas. It effectively extracts from the web sites useful terms and identifies taxonomy information which is essential to ontology schema construction. A prototype system is developed to use this method to help create domain ontology schemas.

Second, this study develops two complemented methods which are light weighted and more semantic web oriented to address the issue of ontology data generation. One method allows users to convert existing structured data (mostly XML data) to ontology data; another enables users to create new ontology data directly with ease.

In addition, a web-based system is developed to allow users to manage the ontology data collaboratively and with customizable security constraints.

Third, this study also proposes two methods to perform ontology data matching for the improvement of ontology data quality when an integration happens. One method uses the clustering approach. It makes use of the relational nature of the ontology data and captures different situations of matching, therefore resulting in an improvement of performance compared with the traditional canopy clustering method. The other method goes further by using a learning mechanism to make the matching more adaptive. New features are developed for training matching classifier by exploring particular characteristics of ontology data. This method also achieves better performance than those with only ordinary features. These matching methods can be used to improve data quality in a peer-to-peer framework which is proposed to integrate available ontology data from different peers.